Analyzing NOAA Storm Damage Data

Project Report

Matthew Tuttle
CSPB
University of Colorado Boulder
CO USA
matu8468@colorado.edu

Lex Bukowski
CSPB
University of Colorado Boulder
CO USA
Alexi.Bukowski@colorado.edu

ABSTRACT

Severe weather events have a profound impact on society and the world. Storms can destroy infrastructure and property, shut down critical systems, and endanger human lives. The National Oceanic and Atmospheric Administration (NOAA) has tracked and documented severe weather events, the damage they cause, and the number of fatalities dating back to 1950. The following project proposal outlines a data mining and analysis exercise that endeavors to answer interesting questions about the NOAA Storm Event data.

The end goal of the project was to enhance our understanding of severe storms and the damage they cause to humans and property by looking at any possible correlations that may exist in the data. This project utilized data mining techniques to discover correlations between the dataset attributes and the severity (damage costs and injury counts) in the massive amount of data that the NOAA has collected over the past 70+ years. The questions that we were looking to answer with our analysis were: Have weather events become more dangerous to people? Has the damage caused by severe weather events increased over time? Have severe weather events increased over time? Is the magnitude of property damage indicative of higher risk of injury to people? Are some areas more prone to severe weather effects?

By implementing methods such as outlier analysis, classification, and regression, these questions were answered, however gaps, inconsistencies, and a limited time frame of standardized data prevent strong predications from the results.

The discovered answers to all of the questions asked above were that no correlation existed between attributes. The regression trends and summaries all had near zero r-squared values and high P values, meaning the attributes were not related. The first three listed questions all have to do with attributes over time, and all three showed no meaningful correlation with the timeframe. The last two questions had to do with attributes correlating to other attributes (damage to injuries, damage to location, injuries to location). The damage attribute showed no correlation to the injury attribute, and this was confirmed with the location analysis as the locations for higher damage from events were geographically different than the locations with high injury counts.

CCS CONCEPTS

• Mathematics of computing • Mathematics of computing ~ Probability and statistics • Mathematics of computing ~ Mathematical analysis • Mathematics of computing ~ Probability and statistics ~ Probabilistic representations • Mathematics of computing ~ Probability and statistics~ Probabilistic inference problems • Mathematics of computing ~ Probability and statistics ~ Probabilistic algorithms

KEYWORDS

Storm events, Storm damage, Property damage, Natural disaster, Weather events, Severe weather, Weather damage, Weather fatalities

ACM Reference format:

Lex Bukowski and Matthew Tuttle. 2023. Analyzing NOAA Storm Damage Data. *4 pages.*

1 Introduction

The interesting questions we decided to analyze the data set for were: Have weather events become more dangerous to people? Has the damage caused by severe weather events increased over time? Have severe weather events increased over time? Is the magnitude of property damage indicative of higher risk of injury to people? Are some areas more prone to severe weather effects?

Answering these questions will allow for a better understanding of the risk of severe weather events and allow for better prediction of the outcomes of severe weather events. Armed with this information, future weather events, even if more severe, can cause less damage to people and property, potentially saving lives and preventing huge expenditures.

The questions surrounding the occurrence of events over time will give insight into possible other causation factors as well as other interesting questions to be explored. Most notably if severe weather events are increasing over time, does global warming and climate change play a role in the occurrence and severity of these events.

The questions involving the location of severe weather events can assist in mediation planning, response planning, and development planning of the areas.

2 Related Work

In conducting background research on this project, topic, and data set, two previous published papers were discovered. They are discussed below.

2.1 Text Mining of NOAA Data

Emma Louise McDaniel analyzed the NOAA dataset previously by implementing a text mining algorithm to uncover the true nature and severity of a severe weather event. The logic supporting this study was that although important for insurance adjustments, monetary cost is not always the most prevalent or usable metric to determine the scale and severity of a storm event. McDaniel summarizes this argument with the following quote from her abstract "The monetary impact of a disaster is not conducive for disaster preparedness planners to build community resilience."[1]

McDaniel's work was subsequently to text mine the narratives given about the severe weather events included in the database. This work proved to be difficult as the natural language was very unstructured. However, that doesn't make the information contained in the narrative any less important. McDaniel's work was a first step toward extracting and classifying key words and terms out of the storm event narratives to determine a relationship between those key terms and a storm's true (not just monetary) effect. Below is a citation of McDaniel's work, which is also included in the references section at the end of this proposal.

SAC '23: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. March 2023 Pages 653 – 656 https://doi.org/10.1145/3555776.3577211

2.2 Comments on Reliability of NOAA's Storm Events Database

Renato P Dos Santos analyzed the reliability, completeness, and high-level validity of the dataset. Dos Santos understood the limitations of validating every input into the database due to the size of the dataset, but took a high-level approach to do a quick analysis on completeness and consistency of the data.

Dos Santos concluded that although there have been standardization efforts on the part of the NOAA to make the database more usable and effective, there are many non-standard event types in the database. Along with those inconsistencies, Dos Santos added that many damage reports are missing and many damage values are incorrect.

Below is a reference to Renato P Dos Santos work, which is also included in the references section at the end of this proposal.

P dos Santos, Renato, Some Comments on the Reliability of NOAA's Storm Events Database (June 22, 2016). Available at SSRN:

https://ssrn.com/abstract=2799273 or http://dx.doi.org/10.2139/ssrn.2799273

3 Data Set

The data set that this project will work with is the "NOAA Storm Events Database." This database is free and open to the public and can be found at the following URL: https://www.ncdc.noaa.gov/stormevents/ftp.jsp.

Summarized but detailed information about the contents of the database can be found at the same URL. The database includes columns that contain information on the details, locations, and fatalities of each storm event, linked by the primary key of the event ID number.

The Details table includes information such as the timing, general location, event type, reported injuries, reported deaths, crop damage, property damage, and the storm narrative (more information is housed in this table, these are just the most relevant fields for this proposal). The Storm Data Location table includes much more detailed location information and the Storm Data Fatality table does the same for the reported fatalities.

It is worth noting that the database has entries from three different phases. From 1950-1954, the database contains only information on tornado damage. From 1955-1995, damage from thunderstorms, wind, and hail was also included. Finally, from 1996 onward, data was collected for a total of 48 event codes. Care will have to be taken to constrain analysis of events over time to the proper time periods based on the available data.

Reporting standards for the data have changed drastically in the lifespan of the data, with major changes occurring at each of the time intervals mentioned above. For a two-year span between 1993 and 1995, events were recorded in unformatted text files. Only tornado, thunderstorm, wind, and hail events were extracted from these files (the same events recorded from 1955 through 1992). The first attempt at meaningful standardization was taken in 1996 with the standardization of event types. As mentioned in the analysis of previous work involving this data set, that standardization still has significant issues, especially do to the important nature of the free text and description fields for classifying these events. Due to these facts, significant work will be done to clean this data set for analysis, and most analysis will be completed on the events post-1996. It is important to note that analysis done on the entire data set will be skewed toward the tornado, wind, thunderstorm, and hail events that have a broader range and timeframe of data collection.

4 Main Techniques Applied

4.1 Preprocessing

There was a significant amount of preprocessing that needed to be completed on this data set. The heart of the dataset was the fields EVENT_TYPE, PROPERTY_DAMAGE, CROP_DAMAGE, INJURIES_DIRECT, and INJURIES_INDIRECT. All of these fields had some significant issues that needed to be addressed before the data could be analyzed.

The EVENT_TYPE field officially has 48 possible event types, but the database includes a total of 70 different labels. Many of these non-standard labels are combinations of existing labels. The non-standard labels were manually recategorized and labels were combined using judgment on event similarity (i.e. snow event was combined with blizzard event).

The PROPERTY_DAMAGE and CROP_DAMAGE fields represent storm damage in dollar value. However, these are not numeric fields but instead are stored as strings with a magnitude signifier as the last character. For example, damage of $25,000 might be stored as '25.0K' with K signifying a magnitude of 1,000. There are a number of non-standard signifiers (e.g. '?') which resulted in the data rows being removed. Additionally, there are a significant number of entries in the data set that have empty values for both damage types. Given that inclusion in the data set is predicated on damage this is likely an error and they were removed from the analyzed data set.

Finally, the INJURIES_DIRECT and INJURIES_INDIRECT fields appear to be much less problematic as they are integer valued. However, only 0.7% of all data entries caused either a direct or indirect injury. Additionally, over half the total injuries in the database were caused by 5 events. With such sparse/concentrated data, conclusions will have to be drawn very carefully.

4.2 Data Integration

Population and infrastructure density are a significant variable in storm damage that is not currently captured in the database. The database does provide standardized region labels for storm locations as well as the latitude and longitude of storms. These fields were used to tie in US census data to analyze the damage and severity of storm events by location and population density.

4.3 Classification and Analysis

After completing the data preprocessing and the added data integration, classification and analysis was performed on the cleaned and complete data set to determine possible correlations that exist in the dataset.

5 Evaluation Methods

5.1 Methods

The possible correlations between attributes of the data set and/or if the occurrence of attributes is increasing (or decreasing) with time were evaluated using various methods. Graphically, information from scatter plots and box plots was used along with more numerical methods of R-squared and Pearson correlation coefficients.

Due to previously referenced work completed by P dos Santos, evaluation of the data set as a whole needs was addressed as well. The data set was analyzed for completeness and consistency while evaluating correlations. Confidence intervals and outlier analysis was utilized to help determine if correlations appear poor due to poor data quality or accurate data analysis.

5.2 Drawing Conclusions

The goal of this project was to make informed conclusions about correlations between severe weather events occurrence, damage, and death toll by analyzing the other factors provided in the data set. The last step of the project was drawing these conclusions from the data processing, integration, and classification steps. Care was taken to not apply unnecessary causation, but simply show determined correlations.

As described in the project proposal, a possible conclusion for this project is that the data set, although encompassing over 70 years of information, does not contain enough structured information to determine interesting correlations with a strong support basis. This was noted as a shortcoming of the dataset in both of the previous works cited above.

6 Tools

A number of tools were used for this analysis. An exploratory analysis of the data was performed with python using pandas and numpy paired with matlibplot and seaborn for data visualization.

Pandas data frame library was used to create a data frame by importing the CSV files containing the years of storm event data. The data frame allowed for data manipulation, cleaning, and visualization using other tools.

Seaborn's plotting abilities were used to help gain an understanding of possible trends in the data set by plotting each event type vs the direct injury count. The seaborn library allowed for the generation of a scatter plot for a visualization of the injury count density for each event type and a box plot to determine the range of meaningful data points for each event type. This initial visualization of the data provided initial insight into the proposed questions and give a starting point for more analysis.

Classification and regression analyses was done in python using a combination of the sklearn and statsmodel libraries.

Documentation for all described libraries is readily available online.

6.1 Exploratory Tools

Pandas and Numpy are libraries built on top of the Python language that will help simplify, clean, and analyze the dataset. Both tools will be used extensively with the goal of standardizing and gaining a high level understanding of the data. This will allow for effective classification and regression analysis, as opposed to choosing data traits randomly. Matlibplot and seaborn will help visualize the information gained during this data scrubbing process and will also provide visuals for final presentation of the data.

6.2 Classification and Regression Tools

The sklearn and statsmodel libraries are tools built on top of python that will allow for effective classification and regression on data traits that were identified as important and relevant from the cleaning phase.

7 Key Results

7.1 Storm Events Over Time

To answer the question on if storm events are becoming more frequent and more severe, the number of events and the two severity factors chosen (damage and injuries) were modeled against time.

7.1.1 Number of Storm Events

The first analysis that was completed was on the number of severe storm events that have occurred each year. This proved to be difficult due to the reporting standards that have changed multiple times throughout the lifespan of the database. Starting in 1950, only tornados were recorded. In 1954 tornado, hail, and wind events were recorded. Starting in 1996, 48 event types were standardized and the database become slightly more structured. Because of the increase in event reporting, it was difficult to data mine the trend of number of events occurring each year. Regression trends show an increasing number of events, but each of those jumps can be tied to a change in reporting standard and is not showing a true pattern or change.

This can be seen visually in Figure 1. There appears to be an increasing trend in the number of events, but that is due to more event types being reported (more colors in each bar).

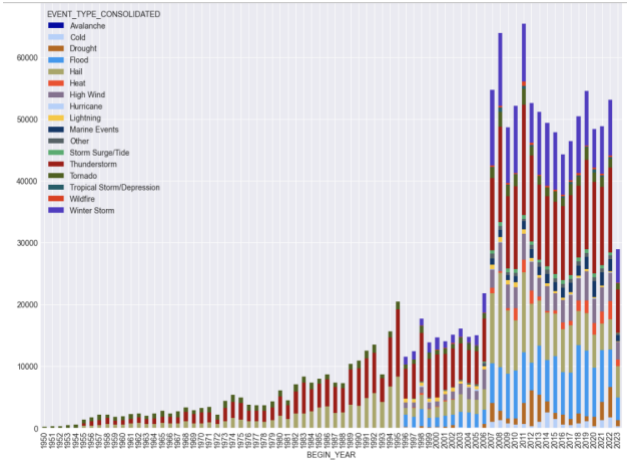

Figure 1: Severe Weather Event Counts by Type and Year

A regression was performed on the data base for events after 2008 (when events such as, tide surges, marine events, and heat events appear to be added to the data set, although not noted by the NOAA, none appeared before then), however the time frame was too short to gain any insight into the data, no statistically relevant trend existed.

The key results from this analysis provided more information on the state of the data set instead of answering the question that we asked. This topic will be discussed in more detail in section 7.3.

7.1.2 Damage

Next, the damage attribute was mined to determine the trend of damage cost over time. For this analysis, the data set was trimmed down to include only the years after 2000 due to the lessons learned from the previous analysis. A regression analysis was completed using statsmodel and is shown below in Figure 2.

OLS Regression Results

Dep. Variable:	DAMAGE_PROPERTY_NUMERIC	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.048
Method:	Least Squares	F-statistic:	1.925e-05
Date:	Wed, 13 Dec 2023	Prob (F-statistic):	0.997
Time:	20:18:00	Log-Likelihood:	-581.27
No. Observations:	23	AIC:	1167.
Df Residuals:	21	BIC:	1169.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.436e+10	1.51e+12	0.009	0.993	-3.14e+12	3.16e+12
END_YEAR	3.303e+06	7.53e+08	0.004	0.997	-1.56e+09	1.57e+09

Omnibus:	24.855	Durbin-Watson:	1.905
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36.310
Skew:	2.216	Prob(JB):	1.30e-08
Kurtosis:	7.271	Cond. No.	6.10e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.1e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 2: Regression Analysis for Damage Cost Over Time

The regression summary shows an R-squared value of 0.000 and a high P value. These values show that there is no correlation between the two attributes of damage cost and time. From this, we can conclude that statistically there is no increase in storm event damage cost in more recent years.

This can be seen visually in Figure 3 which contains a scatter plot of the data with the regression line plotted. The regression line shows no trend either direction.

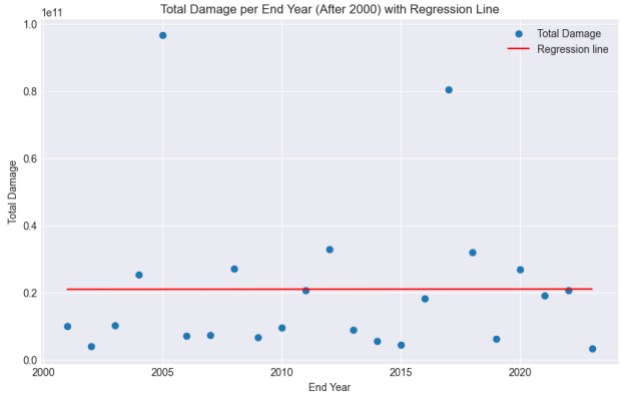

Figure 3: Scatterplot and Regression Line of Damage per Year

7.1.3 Injuries

The other attribute that measures storm event severity is the injury count. The injury count attribute was mined against time as well to determine if there was a pattern or trend.

The same inconsistency issues exist with these attributes in the dataset. In order to work around those issues, the tornado event type was chosen due to it having the largest amount of data available. Choosing a single event type to generalize the injury rate over time eliminates the reporting issues of changing dataset standards.

A similar regression analysis was conducted with the results shown below.

OLS Regression Results

Dep. Variable:	INJURIES_DIRECT	R-squared:	0.068
Model:	OLS	Adj. R-squared:	0.055
Method:	Least Squares	F-statistic:	5.253
Date:	Wed, 13 Dec 2023	Prob (F-statistic):	0.0248
Time:	20:17:55	Log-Likelihood:	-627.29
No. Observations:	74	AIC:	1259.
Df Residuals:	72	BIC:	1263.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.052e+04	1.27e+04	2.396	0.019	5126.008	5.59e+04
END_YEAR	-14.6949	6.412	-2.292	0.025	-27.476	-1.914

Omnibus:	63.549	Durbin-Watson:	1.856
Prob(Omnibus):	0.000	Jarque-Bera (JB):	304.536
Skew:	2.729	Prob(JB):	7.43e-67
Kurtosis:	11.305	Cond. No.	1.85e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.85e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 4: Regression Analysis for Injuries Over Time

This regression summary also has an R-squared value of almost zero (0.068) meaning that there is no correlation between the attributes of injury count and time.

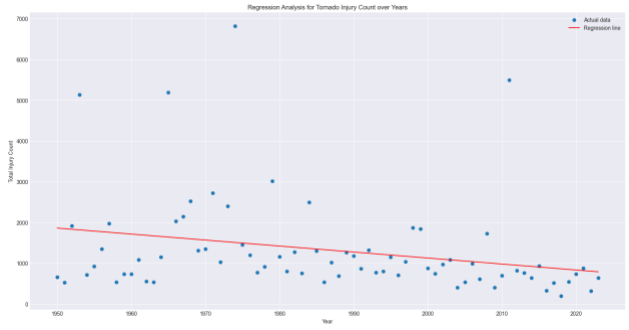

Figure 5: Scatterplot and Regression Line of Injuries per Year

This scatter plot appears to show a slight downward trending regression line; however, the small R-value means it is statistically not relevant and is likely due to the high outliers that can be seen on the scatter plot.

7.1.4 Severity of Extreme Events

One thing that is clear from the data set is that outlier events are the main driver of total property damage. Figure 6 shows the share of property damage each year caused by different quantile events, with the dominant factor showing to be 99th percentile events. Figure 7 breaks this down further, showing the share of total property damage each year caused by different quantiles. This shows well over 80% of all damage each year is caused by 99th percentile events since 2000.

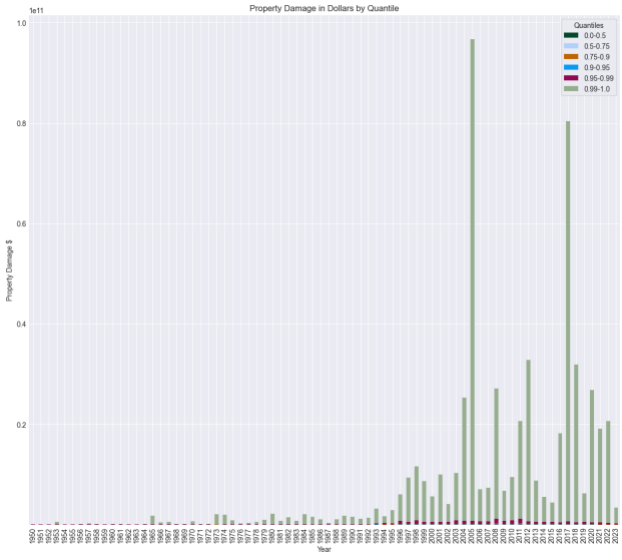

Figure 6: Total Property Damage per Year by Quantile

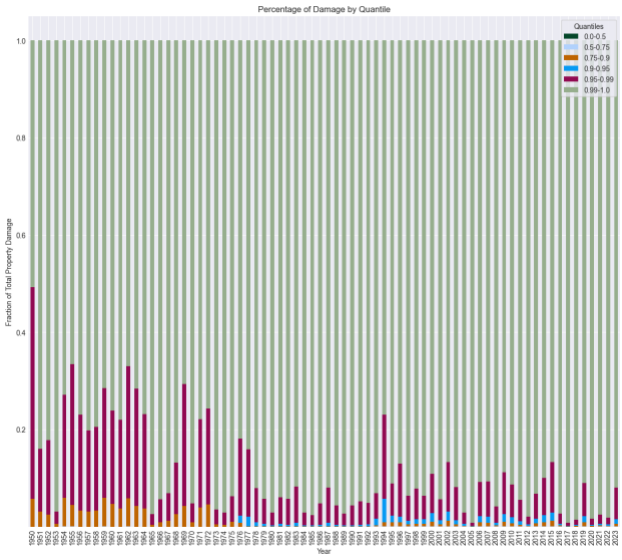

Figure 7: Share of Total Property Damage per Year by Quantile

7.2 **Storm Events by Location**

Given that there are no clear trends of storm damage or injuries changing over time at a nationwide level, further investigation was warranted to see if more localized effects could be discerned by plotting property damage and injuries by location. Summations of property damage and injuries were performed at a county level to give reasonable granularity. As a side effect, however, this analysis excluded larger events that spanned multiple counties or states and were not given singular county codes. The resulting plots are shown in Figures 8 and 9.

Counties with high property damage appear to be concentrated in the gulf coast region, areas with high tornado activity, and in southern California. It also appears that the coastal counties do not have high injury totals to match the high property damage. This may be indicative that widespread flooding events cause fewer injuries than similarly damaging wind events in tornado areas.

It is also worth noting that since these plots include all data from 1950, resulting totals will be skewed towards damage caused by tornadoes.

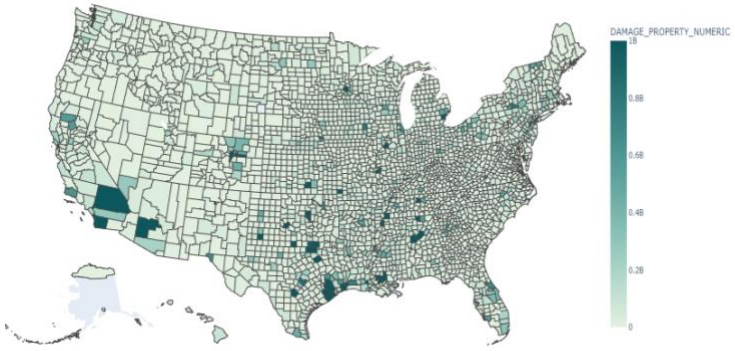

Figure 8: Total Property Damage by County

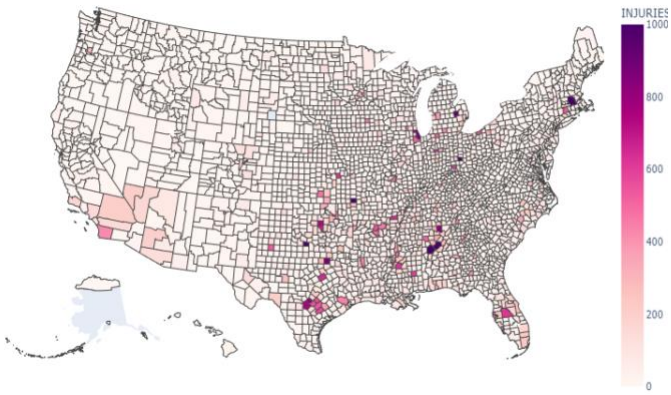

Figure 9: Total Injuries by County

7.3 Correlation Between Damage and Injuries

Another interesting question that was asked was if the magnitude of property damage is indicative of higher risk of injury to people. These damage and injury attributes were pulled from the dataset and a regression analysis was completed. The summary can be seen below in Figure 10.

OLS Regression Results

Dep. Variable:	Damage	R-squared:	0.004
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	5038.
Date:	Thu, 14 Dec 2023	Prob (F-statistic):	0.00
Time:	15:22:18	Log-Likelihood:	-2.3727e+07
No. Observations:	1260201	AIC:	4.745e+07
Df Residuals:	1260199	BIC:	4.745e+07
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.881e+05	3.24e+04	11.977	0.000	3.25e+05	4.52e+05
Injuries	5.285e+05	7445.901	70.981	0.000	5.14e+05	5.43e+05

Omnibus:	6765761.370	Durbin-Watson:	1.848
Prob(Omnibus):	0.000	Jarque-Bera (JB):	595481093955496.250
Skew:	278.789	Prob(JB):	0.00
Kurtosis:	106494.206	Cond. No.	4.35

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 10: Regression Summary for Damage and Injury Count

The very small R-squared value of 0.004 shows that there is no correlation between the amount of damage caused by a storm event and the amount of injuries caused by the same event. This data may also be skewed by the inconsistencies of the data set though, as a large amount of damage costs are related to crop damage caused by tornados, which are the most prevalent storm event in the data set since they have been tracked the longest.

The scatter plot of the data with the regression trend can be seen in Figure 11.

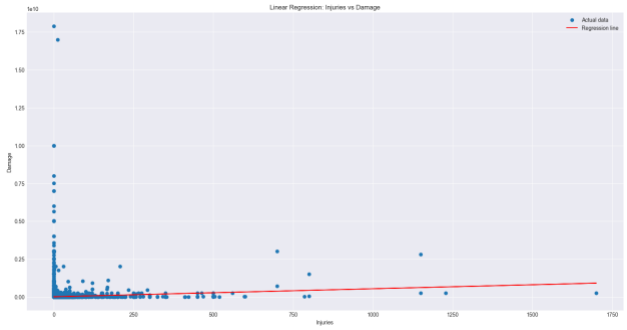

Figure 11: Scatter Plot and Regression Trend of Damage vs Injury Count

The lack of correlation between the damage cost and injury count of the events can also be seen in the location analysis covered in section 7.2. The counties with the highest property damage (Figure 8) being different than the counties with the highest injury counts (Figure 9).

7.4 Data Set Evaluation

Overall, the analysis completed on the data set showed few correlations and trends between attributes. It should be noted, that the inconsistencies and changing in reporting standards are likely what is contributing to this. Significant effort went into data cleaning and preprocessing to standardize and scrub the data set, however, the data cleanliness issue still seems to be seen in the analysis. These issues were echoed in the literature survey completed for this data set.

After analyzing the time and location trends, a recommendation moving forward would be to track storm events at a more local level, where the effects and attributes could be tracked in a more standard and meaningful way.

8 Applications

The NOAA Storm Events Database is an invaluable repository of historical weather data, but it will need to be used judiciously to draw conclusions about future weather patterns. There are few meaningful trends that apply across the United

States geographically and across different types of weather events. Future work will likely be much more fruitful if focused on regional trends and a smaller subsection of weather event types.

ACKNOWLEDGMENTS

Acknowledgements should be made to both Emma Louise McDaniel and Renato P Dos Santos whose previous work with the data was taken into consideration when determining the proposed work and evaluation methods included in this proposal.

Professor Kristy Peterson (University of Colorado Department of Computer Science) and the University of Colorado CSPB 4502 Fall 2023 course participants also provided feedback, critiques, and suggestions to this proposal idea and draft.

REFERENCES

[1]SAC '23: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. March 2023 Pages 653 – 656 https://doi.org/10.1145/3555776.3577211

[2]P dos Santos, Renato, Some Comments on the Reliability of NOAA's Storm Events Database (June 22, 2016). Available at SSRN: https://ssrn.com/abstract=2799273 or http://dx.doi.org/10.2139/ssrn.2799273